

Flem, B., Reimann, C., Fabian, K., Birke, M., Filzmoser, P. and Banks, D. (2018) Graphical statistics to explore the natural and anthropogenic processes influencing the inorganic quality of drinking water, ground water and surface water. *Applied Geochemistry*, 88(Part B), pp. 133-148.(doi:[10.1016/j.apgeochem.2017.09.006](https://doi.org/10.1016/j.apgeochem.2017.09.006))

This is the author's final accepted version.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/147765/>

Deposited on: 13 September 2017

# Graphical statistics to explore the natural and anthropogenic processes influencing the inorganic quality of drinking water, ground water and surface water

Belinda Flem<sup>1</sup>, Clemens Reimann<sup>1</sup>, Karl Fabian<sup>1</sup>, Manfred Birke<sup>2</sup>, Peter Filzmoser<sup>3</sup>, David Banks<sup>4,5</sup>

<sup>1</sup>Geological Survey of Norway, Postboks 6315 Sluppen, 7491 Trondheim, Norway

<sup>2</sup>Federal Institute for Geosciences and Natural Resources, Wilhelmstrasse 25 – 30, 13593 Berlin, Germany

<sup>3</sup>Dept. of Statistics & Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria

<sup>4</sup>School of Engineering, James Watt Building (South), University of Glasgow, Glasgow, G12 8QQ, UK.

<sup>5</sup>Holymoor Consultancy Ltd., 360 Ashgate Road, Chesterfield, Derbyshire, S40 4BW, UK.

## Highlights

- Cumulative distribution function plots are a powerful exploration data analysis tool.
- Despite different origins of waters, median values and ranges are for many elements comparable.
- High concentration of B, Be, Br, Cs, F<sup>-</sup>, Ge, Li, Rb, Te and Zr characterise deeper-seated, mature groundwaters.
- Correlation heatmaps with dendrograms unveils unexpected elemental correlations from plumbing materials.

## Keywords

Cumulative distribution function

Boxplot

Heatmap

Compositional data

European surface water

Groundwater

European bottled water

European tap water

Abstract: Plots of cumulative distribution functions (CDF) are a simple but powerful exploratory data analysis (EDA) tool to evaluate and compare statistical data distributions. Here, empirical CDF plots are used to compare results of four large (476 to 884 samples) national- to continental-scale inorganic water chemistry data sets: (1) European surface water, (2) European tap water, (3) European bottled waters as a proxy for groundwater and (4) Norwegian crystalline bedrock rock groundwater, all analysed at the same laboratory, albeit at different times. For many parameters (e.g., Ba, Cl<sup>-</sup>, K, SO<sub>4</sub><sup>2-</sup>) median values and ranges are, given the differing origins and, in some cases, treatment processes of the waters, surprisingly comparable. Unusually high concentrations of some other elements (e.g., B, Be, Br, Cs, F<sup>-</sup>, Ge, Li, Rb, Te and Zr) appear to be characteristic of deeper-seated, mature groundwaters. Other influences that can be inferred include contamination from well construction or plumbing materials (Cu, Pb, Zn – in tap waters, bottled waters and Norwegian groundwaters), water treatment (Fe, Mn – in tap- and Norwegian groundwater), bottle materials (Sb - bottled waters). The empirical CDF plots also reveal analytical issues for some elements (excessive rounding, element interferences). The best reference for natural and uncontaminated 'water' is probably provided by the mineral water samples, representing 'deep groundwater' at the European scale.

## 1. Introduction

The chemical “fingerprint” of ground- and surface waters has been used to study natural and anthropogenic impacts on water quality in several contexts including: water circulation (e.g., Dragon and Marciniak, 2010), catchment conceptual modelling (e.g., Banks, 2014), tracing of origin (e.g., Dragon and Gorski, 2015), contamination (e.g., Luís et al., 2011), mineral exploration (e.g. Wang et al., 1995), national baseline characterisation of groundwater bodies (Shand et al., 2007; O'Dochartaigh et al. 2015) and human health (e.g., Bhowmik et al.,

2015). The majority of studies are at a local to national scale, however, although continental-scale studies have been published (e.g., mineral waters in the Tracing the Origin of Food – TRACE-project; Bertoldi et al., 2011). The problem with most of these studies is that the sample sets cover different parts of Europe and were analysed in different laboratories with varying methods for differing suites of elements. They are thus difficult to compare and are not necessarily suited for gaining an overall impression of “water quality” at a European scale. To date, the most comprehensive datasets at a European scale have been provided by the Geological Surveys of Europe (formerly FOREGS – Forum of European Geological Surveys, now EuroGeoSurveys) for surface water (Salminen et al., 2005) and groundwater (Reimann and Birke, 2010).

Concentrations of chemical elements in bedrock, soil, sediments, plants and water are reported in relative units like wt%, mg/kg or  $\mu\text{g/L}$ . These units indicate and imply that earth and environmental scientists are dealing with compositional data (CoDa) in practically all their investigations. The important consequences of CoDa for statistical data analysis have first been discussed by Aitchison (1986). In short, CoDa values have a finite range (e.g., 0%-100%, or 0- $10^6$  mg/kg), and also the sum of all variables is less or equal to the maximum value. Mathematically this implies that CoDa vectors do not belong to a linear infinite Euclidean space, but are constrained to the Aitchison simplex. Because most standard statistical techniques are based on Euclidean spaces, and Euclidean distances between data vectors, their application to CoDa are outright false.

For water data it has sometimes been argued that the error involved is negligible because of the very low concentrations of all elements (with the exception of H and O) in water samples; however, others have argued that water chemistry data should also be treated statistically as compositional data (Buccianti and Pawlowsky-Glahn, 2005; Engle and Rowan, 2014; Engle et al., 2014).

Notably, some important statistical quantities still have a well-defined meaning for CoDa. These include individual concentration mean values and all concentration quantiles. For a set of samples ( $S_1, \dots, S_N$ ) and some element concentrations ( $c_1, \dots, c_N$ ), the value  $c_i$  in sample  $S_i$  denote the amount of mass  $m_i$  of the element per reference mass  $M$  or volume:

$$c_i = m_i / M$$

This is an important measurable quantity for geochemists, e.g. to determine toxicity, or ore tonnage. Its quantiles in robust statistics are well-defined measurable quantities with clear physical meaning. Also the mean value:

$$\bar{c} = \frac{c_1 + \dots + c_N}{N} = \frac{m_1 + \dots + m_N}{NM}$$

still has a clear physical meaning as the concentration of the element if all samples,  $S_1, \dots, S_N$ , are well mixed to one sample. However, for example, the standard deviation of these concentrations has lost its mathematical background (Euclidean distance or normal distribution), because the concept of standard deviation is implicitly based on the false assumption that concentrations can take arbitrary real values. For this reason most classical statistical test methods do not apply to geochemical data.

Well recognised log-ratio transformations for geochemical data, such as the *clr* (centered log-ratio) or *ilr* (isometric log-ratio) transformations (see, e.g., Pawlowsky-Glahn and Buccianti, 2011) enable the geochemists to work in the usual Euclidean geometry for which most statistical methods are designed for. The original concentration data ( $c_1, \dots, c_N$ ) are transformed to values that are related to all other element concentrations. These values contain no isolated information about the element alone, and they could not be used for e.g., studying toxicity thresholds or calculating ore tonnage. Since these values contain information concerning all other elements in the given composition, they can yield a very

different picture than the absolute concentration data alone. This is an incentive to use both sources of information for the exploratory data analysis tools presented later on.

Note that the numerical ranking between concentrations and correlated values can be different. This implies that quantiles and rank statistics of individual concentrations and correlated values of the same element can be different. Yet, both are mathematically meaningful, they only describe conceptually different quantities.

In the initial stages of data analysis, geochemists are usually interested in element concentrations, in the median (or mean), in certain percentiles of the data distribution and in the variation within the data. All these are directly related to the concentration of the studied element in their sample material. Action levels or guideline values are all provided in concentration-related units. When interested in absolute concentration data it is thus wholly justified to use several of the classical statistical methods on the untransformed (or log-transformed) data. The original data can, for example, be used to compare between the median, certain percentiles or the total variation observed for different datasets.

In earlier studies, graphical exploratory data analysis techniques have been arguably somewhat under-utilised, although the British Geological Survey's "Baseline chemistry of groundwater in UK aquifers" programme made extensive use of CDF diagrams (Shand et al., 2007). In this paper, we aim to demonstrate the utility of some very simple graphical techniques (empirical cumulative distribution function ECDF plots, boxplots) to explore large national or continental-scale hydrochemical data sets. In a second step, correlation analysis is performed using the correctly log-ratio transformed data which consider the compositional nature of the analytical results. This paper demonstrates a number of simple tools that can be used to investigate hydrochemical data, utilising both the absolute data and the log-ratio transformed data.

## 2. Methods

### 2.1 Sampling and analysis

The data sets presented in this paper originate from the following projects:

(1) European surface waters. The Forum of European Geological Surveys (FOREGS) “Geochemical atlas of Europe” project (described in detail by Salminen et al., 2005) collected 807 stream water samples from second order drainage basins across Europe. This dataset is hereafter referred to with the abbreviation SW.

(2) European tap water: the EuroGeoSurveys European Groundwater Geochemistry (EGG) Project collected 579 samples across Europe and organised their analysis as described in Reimann and Birke (2010), Banks et al., (2015) and Flem et al., (2015). The abbreviation TW will be used for this dataset.

(3) European bottled water. 1785 bottled water samples were collected from shops and supermarkets during the EuroGeoSurveys European Groundwater Geochemistry (EGG) Project. These bottled waters were typically what would be classified as “spring” or “mineral” waters and it was intended that these could be regarded as a “proxy” for relatively uncontaminated European groundwaters. The source locations of the waters were identified, (Reimann and Birke, 2010) and the results reported here represent 884 unique locations. The abbreviation BW will be used for this dataset.

(4) Norwegian crystalline bedrock groundwater: A subset of 476 samples was carefully selected, so as to be lithologically and geographically as representative as possible, from a larger set of around 1600 groundwater samples from wells and boreholes (mostly small, private sources) drilled into Precambrian or Palaeozoic crystalline bedrock in Norway (mostly southern Norway). The study was organised by the Geological Survey of Norway

and is described by Banks et al. (1998) and Frengstad et al. (2000, 2001). It is recognised that Norwegian crystalline bedrock groundwater has a rather distinctive hydrogeochemistry and is not representative of broader groundwater quality at a European scale. The abbreviation GW will be used for this dataset.

Of the sample sets described above, only the European surface waters were subject to field filtration (at 0.45  $\mu\text{m}$ , Schleicher & Schuell pyrogen free); samples in the other sets were sampled “as found”, although in the Norwegian bedrock groundwater data set, efforts were made to collect samples as near to the well-head as possible and prior to any water treatment. Sampling and analytical procedures are described in detail in the publications cited above. Some of the accompanying parameters in the GW data set; pH, alkalinity,  $\text{Cl}^-$ ,  $\text{F}^-$ ,  $\text{NO}_3^-$  and  $\text{SO}_4^-$  (by ion chromatography, IC) and some major elements K, Na, Mg, Ca, Si, Fe and Mn (by inductively coupled plasma atomic emission spectrometry, ICP-AES), were analysed at the Geological Survey of Norway. The IC measurements of  $\text{Cl}^-$ ,  $\text{F}^-$ ,  $\text{NO}_3^-$  and  $\text{SO}_4^-$  in SW were performed at BGS (British Geological survey, UK). All samples were subject to multi-element analysis by ICP-QMS (inductively coupled plasma quadrupole mass spectrometry), Ag, Al, As, B, Ba, Be, Bi, Cd, Ce, Co, Cr, Cs, Cu, Dy, Er, Eu, Fe, Ga, Gd, Ge, Hf, Ho, La, Li, Lu, Mn, Mo, Nb, Nd, Ni, Pb, Pr, Rb, Sb, Sc, Se, Sm, Sn, Ta, Tb, Te, Th, Ti, Tl, Tm, U, V, W, Y, Yb, Zn and Zr or ICP-AES Ca, K, Mg, Na, P, Si, Sr (except the above mentioned analysis of GW) at the laboratory of the Federal Institute for Geosciences and Natural Resources (BGR) in Hannover, Germany, within a time period of approximately ten years. These four data sets were measured by three different equipment generations: (1) Elan Perkin Elmer PE-Sciex 250, (2) Elan Perkin Elmer PE-Sciex 5000 and (3) Agilent 7500 ce. The first two ICP-QMS used for measuring of the Norwegian groundwater and surface water, worked without collision cells. Due to the long time range for the different studies and the use of different



generations of ICP-QMS machines, detection limits do vary between the data sets. Quality control was extensive for each of the projects and is documented in detail in the publications cited above.

## 2.2 Data Analysis

### 2.2.1 Cumulative distribution plots (ECDF)

One of the simplest and most powerful techniques for examining a number of large data sets is some form of cumulative probability (CP) diagram, where the data are sorted from low to high values. The family of CP diagrams (e.g., Tennant and White, 1959, Sinclair, 1974, 1991, Stanley and Sinclair, 1987), the ECDF-plot (ECDF = empirical cumulative distribution function) is constructed around a discrete step function that jumps with each data value by  $1/n$ , where  $n$  is the number of data points (see e.g., Reimann et al., 2008). The graphic permits an easy inspection of the central portion of the data distribution, which is often the most interesting section for hydrochemists as it characterises the 'typical' composition of the water, but which is also the most difficult to inspect in other more conventional plots, if much of the data clusters in this area. However, the ECDF plot equally presents the low and high 'tails' of the distribution, which could yield important information about, e.g., contamination or natural processes that give rise to extreme concentrations of an element. The y-axis shows the probabilities of the empirical cumulative distribution function between 0 and 1 (this can be presented on a linear scale, or as a normalised probability scale), while the x-axis shows the concentration. As argued above an ECDF plot can be presented for the untransformed (absolute) concentration data, but it could also be presented for the *ilr*-transformed data (Filzmoser et al., 2009c). In the latter case, several important factors for the interpretation based on absolute concentrations are lost, such as the direct link to the element concentrations

(and potential toxicity) and the related percentiles, the observation of detection limit issues and the detection of outliers at both ends of the distribution. Thus, in this paper only the classical descriptive ECDF-plots will be shown. However, to allow the reader to directly compare these concentration based diagrams with similar diagrams (where the data are put into the correct geometry) for the *ilr*-transformed variables are made available for all elements in the datasets in the supplementary material (Fig. S1). Differences in the distribution between the four datasets are for most elements rather small comparing the ECDF-plot of absolute concentrations and ECDF-plots of *ilr*-transformed data. One effect of the *ilr*-transformation is that all distributions look rather smooth, the many detection limits issues appear to be gone (see, for example, Ag in the supplementary material). The reason for this is that concentrations below the detection limit which have been set to a constant value of half the detection limit is divided by the geometric mean of all other variables for each sample the *ilr*-transformed plots and thus ‘values’ appear for all samples. At first glance this might give the impression that the *ilr*-transformation has ‘improved’ the data-set’s distribution. That is of course not the case, as the transformation is still applied to an essentially arbitrary value (half the detection limit).

Comparison of several datasets in one ECDF-plot readily highlights similarities and differences between the elemental distributions. The ECDF-plot is not based on any assumptions regarding the underlying data distribution; the graphic is thus ideally suited as an exploratory data analysis tool for large datasets in the first stages of the data analysis.

### 2.2.2 Boxplots

Boxplots, as introduced by Tukey (1977) as a “summary plot” of the data distribution of an element, provide in principle the same information as ECDF-plots. The visualisation of the

data is, however, in some cases more immediate because of the focus on percentiles and the detection of outliers. The box contains the central 50% of the data and represents the interquartile range (IQR), with the upper and lower edges of the box corresponding to the 75th and 25th percentiles of the data set. The ‘inner fence’ (a boundary beyond which individual data points are considered extreme values or possible ‘outliers’) is defined as the box extended by 1.5 times the length of the box towards the maximum and minimum (Reimann et al., 2008). Graphically this ‘outlier boundary’ is presented by the whiskers, which are at both ends plotted to the furthest observation inside the inner fence. The boxplot thus indicates at a single glance (1) whether the data distribution is symmetrical or (2) skewed, (3) the location and value of the median, (4) whether outliers exist or not and (5) how far the outliers are removed from the main body of data. Again, due to the compositional nature of the data, it could be interesting to compare the concentration-related boxplots of the original (or simply log-transformed concentration data) to box plots of the *ilr*-transformed data. These plots are provided in the supplementary material for comparison (Fig. S1). Again, changes are most often quite subtle. When performing EDA, the scale chosen for the concentration axis can significantly affect the graphical appearance of both ECDF-plots and boxplots; it is therefore recommended to ‘experiment’ with the scaling of the axis during the early stages of the data analysis.

The R library "StatDA" (Filzmoser, 2015) has been used for preparing both, ECDF plots and boxplots.

### 2.2.3 Bivariate, Multivariate and Component Analysis

Compositional data consist of real-valued vectors  $x=(x_1, \dots, x_D)^t$  with  $D$  strictly positive components describing the parts on a whole, and which carry relative information (Aitchison,

1986; Egozcue, 2003). Thus the correlation structure of compositional data is strongly spurious and results of many multivariate techniques become doubtful without a proper transformation of the data (Filzmoser et al., 2009a)

For multivariate data analysis, when the aim is to detect connection between the variables of a given composition, appropriate data transformations (the family of log-ratio transformations) for compositional data (e.g., centred log ratio *clr* and/or isometric log ratio *ilr* transformations; Filzmoser et al., 2009b, Pawlowsky-Glahn and Buccianti, 2011) are usually required. This starts as early as at the apparently straightforward correlation analysis stage. In this paper, the recently developed version (Kynclova et al., 2017, or see a worked example in Reimann et al., 2017) of a correlation analysis for compositional data has been used. This method considers all relative information of two variables (compositional parts) to the remaining variables, and constructs orthonormal coordinates, for which standard correlations can be computed. This can be done for all pairs of variables, and the information can be represented in so-called “heatmaps”, which rearrange the variables according to their similarity, using a hierarchical cluster analysis based on a Euclidean geometry.

The R library ‘robCompositions’ (Templ et al., 2011) was used for the centred log-ratio (*clr*) transformation of the data, while the R libraries ‘Matrix’ (Bates and Maechler, 2016) and ‘gplots’ (Warnes et al. 2016) were applied for the construction of the heatmaps in R version 3.3.1 (R Core Team, 2016).

### 3. Results and discussion

An overview of the data, providing the minimum, median and maximum value of each measured element is given in Table 1. Note that different detection limits may apply for the different data sets.

Table 1. Summary of analysed constituents, minimum (Min), median (Med) and maximum (Max) measured value of each parameter (N=number of samples, NA=not analysed, RA=Rejected analysis). EC is electrical conductivity, tAlk is total alkalinity determined to a titration end-point pH of c. 4.3. Ion chromatography (IC) was used to determine Cl<sup>-</sup>, F<sup>-</sup>, NH<sub>4</sub><sup>+</sup>, NO<sub>3</sub><sup>-</sup>, and SO<sub>4</sub><sup>2-</sup>, pH determined by potentiometry. All other elements were analysed by inductively coupled plasma mass or emission spectroscopy (see text for details).

		Bottled water (N=884)			Tap water (N=579)			Ground water (N=476)			Surface water (N=807)		
	unit	Min	Med	Max	Min	Med	Max	Min	Med	Max	Min	Med	Max
Ag	µg/L	< 0.002	< 0.002	112	< 0.002	< 0.002	5.55	< 0.002	< 0.002	0.0340	< 0.002	< 0.002	0.42
Al	µg/L	< 0.5	1.19	966	< 0.5	2.47	250	< 1	13.5	3630	0.700	17.7	3370
As	µg/L	< 0.03	0.235	89.8	< 0.03	0.190	71.9	0.011	0.18	18.6	< 0.01	0.630	27.3
B	µg/L	< 2	39.2	120000	< 2	15.5	1170	0.75	13.6	452	0.1	15.6	3030
Ba	µg/L	0.05	29.2	26800	0.05	30.1	1660	0.025	15.1	384	0.20	24.9	436
Be	µg/L	< 0.01	< 0.01	64.1	< 0.01	< 0.01	0.18	< 0.005	0.012	6.64	< 0.005	0.009	2.72
Bi	µg/L	< 0.005	< 0.005	0.692	< 0.005	< 0.005	0.094	< 0.001	< 0.001	3.19	< 0.002	0.002	0.160
Br	µg/L	< 3	35.0	21700	< 3	11.0	2070	1.48	30.4	4030	< 10	< 10	7900
Ca	mg/L	0.434	65.9	611	1.20	59.5	157	0.0229	26.9	200	0.226	40.2	592
Cd	µg/L	< 0.003	0.0032	1.13	< 0.003	0.0083	1.43	< 0.002	0.0170	8.09	< 0.002	0.010	1.25
Ce	µg/L	< 0.001	< 0.001	6.16	< 0.001	0.0018	0.740	< 0.002	0.110	27.9	< 0.002	0.055	36.0
Cl <sup>-</sup>	mg/L	0.18	13.4	3630	0.11	14.1	458	< 0.1	9.40	344	0.14	8.76	4560
Co	µg/L	< 0.01	0.0232	16.4	< 0.01	0.0231	2.26	< 0.005	0.0650	37.4	0.01	0.16	15.7
Cr	µg/L	< 0.2	< 0.2	27.2	< 0.2	< 0.2	17.7	< 0.01	0.14	8.94	< 0.01	0.38	43.0
Cs	µg/L	< 0.002	0.0394	415	< 0.002	0.0075	5.19	< 0.001	0.096	19.4	< 0.002	0.006	24.3
Cu	µg/L	< 0.1	0.273	99.7	< 0.1	5.65	1630	0.31	16.3	496	0.08	0.88	14.6
Dy	µg/L	< 0.001	0.00119	0.389	< 0.001	0.00104	0.448	< 0.002	0.022	1.53	< 0.002	0.008	3.43
EC	µS/cm	18.0	589	26500	20.0	365	1810	NA	NA	NA	5.00	300	17100
Er	µg/L	< 0.001	< 0.001	0.773	< 0.001	< 0.001	0.223	< 0.002	0.015	0.69	< 0.002	0.006	2.08
Eu	µg/L	< 0.001	< 0.001	0.447	RA	RA	RA	< 0.002	0.003	0.400	< 0.002	0.0047	0.87
F <sup>-</sup>	mg/L	< 0.003	0.188	10.7	< 0.003	0.087	1.45	< 0.05	0.212	5.89	< 0.05	0.100	1.55
Fe	µg/L	< 0.5	0.687	13500	< 0.5	3.21	1290	0.500	33.6	8590	< 1	67.0	4820
Ga	µg/L	< 0.005	< 0.005	3.88	RA	RA	RA	< 0.002	0.0130	2.63	< 0.002	0.0110	0.170
Gd	µg/L	< 0.002	< 0.002	0.662	< 0.002	< 0.002	0.624	< 0.002	0.024	2.33	< 0.002	0.0100	4.32
Ge	µg/L	< 0.03	< 0.03	110	< 0.03	< 0.03	1.02	< 0.002	0.0170	1.45	< 0.005	0.009	0.440
Hf	µg/L	< 0.002	< 0.002	1.57	< 0.002	< 0.002	0.0422	< 0.002	0.00400	0.190	< 0.002	0.004	0.120
Ho	µg/L	< 0.001	< 0.001	0.122	< 0.001	< 0.001	0.0826	< 0.001	0.00500	0.260	< 0.002	0.002	0.710
I	µg/L	< 0.2	4.78	4030	0.320	3.23	294	< 0.1	0.600	37.6	RA	RA	RA
K	mg/L	< 0.1	2.10	558	< 0.1	1.60	30.2	< 0.5	2.26	24.3	< 0.01	1.60	182
La	µg/L	< 0.001	0.0023	10.0	< 0.001	0.0023	0.877	< 0.002	0.100	19.4	< 0.002	0.034	16.0
Li	µg/L	< 0.2	10.0	9860	< 0.2	2.65	74.9	0.047	2.90	184	0.005	2.10	356
Lu	µg/L	< 0.001	< 0.001	0.411	< 0.001	< 0.001	0.0239	< 0.001	0.003	0.110	< 0.002	< 0.002	0.300
Mg	mg/L	< 0.01	16.5	4010	0.135	9.61	60.4	0.0250	3.43	26.2	0.048	6.01	230
Mn	µg/L	< 0.1	0.536	1870	< 0.1	0.544	83.1	0.05	15.7	3760	< 0.1	15.9	3010
Mo	µg/L	< 0.02	0.284	74.1	< 0.02	0.233	13.2	< 0.002	1.40	96.0	0.005	0.220	16.0
Na	mg/L	0.4	15.5	8160	0.1	9.47	363	1.08	11.3	285	0.231	6.50	4030
Nb	µg/L	< 0.01	< 0.01	0.537	< 0.01	< 0.01	0.028	< 0.002	0.004	0.29	< 0.002	0.004	0.34
Nd	µg/L	< 0.001	0.00214	5.12	< 0.001	0.00223	2.51	< 0.002	0.120	16.7	< 0.005	0.0400	19.8

NH <sub>4</sub> <sup>+</sup>	mg/L	< 0.005	< 0.005	59.7	< 0.005	< 0.005	1.15	NA	NA	NA	NA	NA	NA
Ni	µg/L	< 0.02	0.178	94.6	0.027	0.381	27.4	0.023	0.53	391	0.03	1.91	24.6
NO <sub>3</sub> <sup>-</sup>	mg/L	< 1	1.32	995	< 1	3.88	44.8	< 0.05	0.669	70.4	< 0.04	2.80	107
P	µg/L	< 6.5	32.6	2860	< 6.5	9.78	6030	< 100	< 100	994	NA	NA	NA
Pb	µg/L	< 0.01	0.0158	2.29	< 0.01	0.118	87.0	0.011	0.36	26.4	< 0.005	0.092	10.6
pH		3.95	6.80	9.90	6.08	7.67	8.63	6.18	8.08	9.58	2.20	7.70	9.80
Pr	µg/L	< 0.001	< 0.001	1.54	< 0.001	< 0.001	0.407	< 0.002	0.027	4.30	< 0.002	0.009	4.70
Rb	µg/L	0.0146	2.12	631	0.045	0.909	44.4	0.028	2.60	32.9	0.09	1.32	112
Sb	µg/L	< 0.01	0.272	4.43	< 0.01	0.0673	3.16	< 0.002	0.0325	8.00	0.005	0.07	2.91
Sc	µg/L	RA	RA	RA	RA	RA	RA	RA	RA	RA	NA	NA	NA
Se	µg/L	< 0.02	0.0544	371	< 0.02	0.115	4.58	< 0.01	0.200	21.1	< 0.01	0.340	15.0
Si	mg/L	0.421	6.50	58.9	0.0935	4.30	36.9	0.771	4.73	14.9	0.0476	3.82	34.3
Sm	µg/L	< 0.001	0.00127	0.671	< 0.001	0.00146	0.658	< 0.002	0.022	2.56	< 0.002	0.009	3.82
Sn	µg/L	< 0.02	< 0.02	1.81	< 0.02	< 0.02	0.247	< 0.002	0.008	45.8	NA	NA	NA
SO <sub>4</sub> <sup>2-</sup>	mg/L	0.01	20.0	20300	< 0.01	26.9	267	2.35	11.9	392	< 0.3	16.1	2420
Sr	µg/L	2.00	326	25500	5.00	177	2850	0.2	144	6340	1.00	109	13600
Ta	µg/L	< 0.005	< 0.005	0.0374	< 0.005	< 0.005	0.0191	< 0.002	0.002	0.037	< 0.002	< 0.002	0.12
tAlk	meq/L	< 0.03	4.70	264	0.102	3.13	13.9	0.0485	1.92	8.06	0.005	2.07	29.6
Tb	µg/L	< 0.001	< 0.001	0.077	< 0.001	< 0.001	0.0874	< 0.001	0.003	0.320	< 0.002	0.002	0.590
Te	µg/L	< 0.03	< 0.03	0.316	< 0.03	< 0.03	0.0372	< 0.005	< 0.005	0.075	< 0.005	< 0.005	0.110
Th	µg/L	< 0.001	< 0.001	0.146	< 0.001	< 0.001	0.0331	< 0.001	0.006	3.06	< 0.002	0.009	0.370
Ti	µg/L	< 0.08	< 0.08	6.34	< 0.08	0.0867	2.02	< 0.01	0.59	495	< 0.01	0.90	16.8
Tl	µg/L	< 0.002	0.00411	2.19	< 0.002	0.00368	1.12	< 0.002	0.00700	0.250	< 0.002	0.005	0.220
Tm	µg/L	< 0.001	< 0.001	0.191	< 0.001	< 0.001	0.0280	< 0.001	0.002	0.0930	< 0.002	< 0.002	0.280
U	µg/L	< 0.001	0.228	229	0.001	0.307	56.2	< 0.001	2.48	749	< 0.002	0.320	21.4
V	µg/L	< 0.1	0.168	48.9	< 0.1	0.174	13.7	< 0.01	0.235	13.9	< 0.05	0.46	19.5
W	µg/L	< 0.05	< 0.05	28.5	< 0.05	< 0.05	64.0	< 0.002	0.0705	66.3	< 0.002	0.007	3.47
Y	µg/L	< 0.001	0.0123	3.49	< 0.001	0.0099	2.70	0.002	0.210	8.13	0.003	0.064	26.6
Yb	µg/L	< 0.001	< 0.001	1.84	< 0.001	0.0011	0.165	< 0.002	0.013	0.63	< 0.002	0.0057	1.79
Zn	µg/L	< 0.2	0.894	651	< 0.2	23.5	5040	0.570	13.9	3610	0.0900	2.65	310
Zr	µg/L	< 0.001	0.0075	165	< 0.001	0.0095	2.08	< 0.002	0.018	7.74	< 0.002	0.053	2.41

### 3.1 Comparison of the four water types: similarities

The ECDF-plots in Fig.1 of Ba, Cl<sup>-</sup>, K and SO<sub>4</sub><sup>2-</sup> demonstrate that despite the different origins of SW (surface water) and BW (ground water), these parameters show a surprisingly similar concentration distribution at a European scale. At the upper end of the data distribution,

however, the bottled waters exhibit greater concentrations of the parameters, however, as would be expected, as groundwater is usually subject to a greater influence of water-rock interaction and evapotranspirative concentration (during recharge) than surface water. ECDF-plots and boxplots, both with absolute concentrations and the proportional *ilr*-transformed values, for all elements measured in all four data sets (Tab.1) are provided in the supplementary material (Fig S1) for comparison.

While the ECDF plot presents all the data in a data set, the boxplot provides a visually simpler summary of the same data. Although the same elements are presented in Fig. 2, as in Fig. 1, the graphical impression is quite different (boxplots of all elements in Tab.1 are provided in the supplementary material, Fig. S1). As the boxplots in Fig. 2 indicate, the main part of the data of Ba, Cl<sup>-</sup>, K and SO<sub>4</sub><sup>2-</sup> are distributed over the same concentration range. The suite of parameters in Fig. 1, in addition to B, Ca, Na, Si, Sr and tAlk, show an overall variation in median concentration between all datasets of less than 60%. If the relative difference between concentrations in BW and TW is calculated from Tab. 1 more than 20 constituents (As, Ba, Ca, Cl<sup>-</sup>, Co, Dy, I, K, La, Mg, Mn, Mo, Na, Nd, Sc, Si, Sm, SO<sub>4</sub><sup>2-</sup>, Sr, Tl, U, V, Y and Zr) exhibit a relative variation in median concentration of less than 50%.

At a European scale, the chemical signatures of the different water types are thus surprisingly similar for many elements. At a more local scale the chemical differences between groundwater and surface water may be more apparent. For example Cao et al. (2016), comparing hydrochemical characteristics of rivers and groundwaters in the Sanjiang plain, China, demonstrated that the median concentration of elements such as K (11% relatively lower in SW) and Cl<sup>-</sup> (26% relatively higher in GW) are quite similar. In contrast, An et al. (2014) demonstrated large discrepancies between K and Cl<sup>-</sup> concentrations in surface water

and groundwaters in the Mekong Delta, Vietnam, probably due to substantial seawater intrusion in rivers.

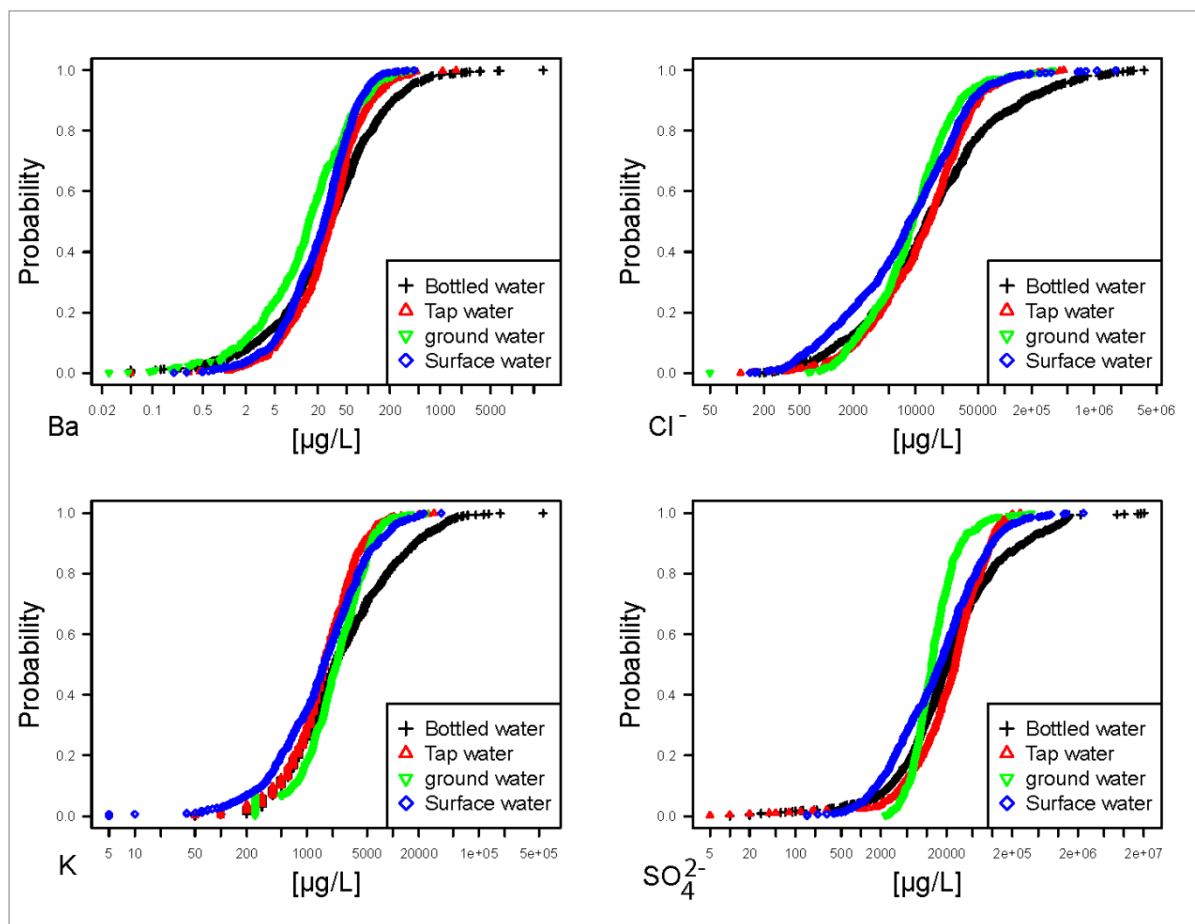


Fig. 1 Empirical cumulative distribution function (ECDF-plot) of Ba, Cl<sup>-</sup>, K and SO<sub>4</sub><sup>2-</sup> in European bottled water (BW), European tap water (TW), Norwegian hard rock groundwater (GW) and European surface water (SW). Note the logarithmic scale of the x-axis of the plot.



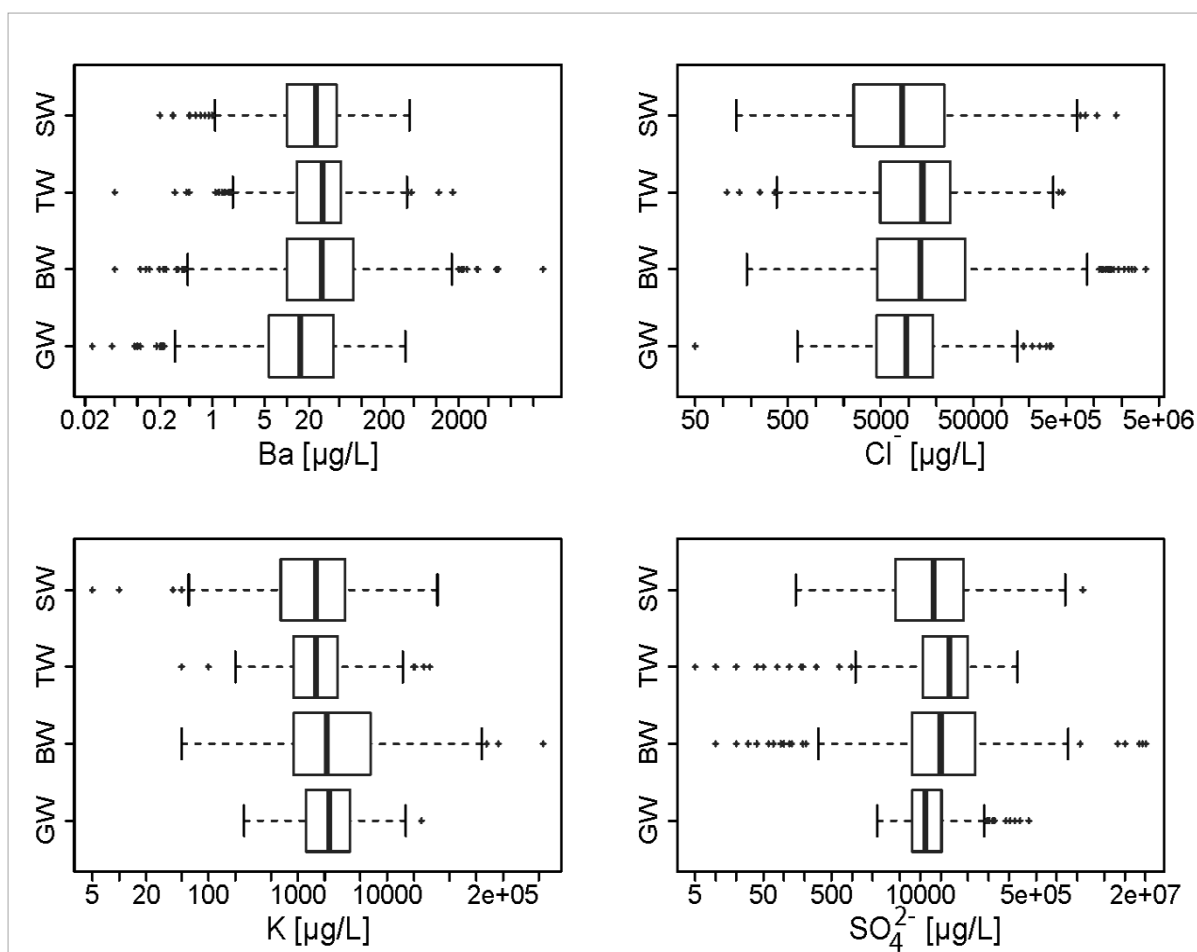


Fig. 2 Boxplots of Ba, Cl<sup>-</sup>, K and SO<sub>4</sub><sup>2-</sup> in European bottled water (BW), European tap water (TW), Norwegian hardrock groundwater (GW) and European surface water (SW). Note the logarithmic scale of the x-axis of the plot.

### 3.2 Comparison of the four water types: Special features

Unusually high concentrations of elements such as B, Cs, Li Na, Sr and Rb (Fig. 3, 4) are associated with bottled water, especially at the upper end of the distribution. This may represent a genuine characteristic of the (often relatively deep, with long residence time) groundwater sources used to supply bottled waters, as the elements named above are typically highly soluble with few solubility ceiling controls, and are often used as indicators of hydrochemical maturity (Banks et al., 2014). However, there may also be purely aesthetic factors to consider: the distributions may simply represent the observation that there is a sector of the European bottled water market (especially in Eastern Europe) which exhibits a

consumer preference for more saline, mineral-rich waters (Banks et al, 2015; Flem et al., 2015). Aside from the generally elevated concentrations in the data set BW, the alkali metals, K, Li, Na, and the metalloid, B, all exhibit similar distributions in the data sets TW, SW and GW. The TW dataset consists of a mixture of both surface water and ground water, which causes the TW to take a intermediate position in these plots. The boxplots for Cs and Rb demonstrate higher median concentrations for bottled (BW) and Norwegian ground water (GW) compared to TW and SW. There is, however, a very wide range in the concentrations within each data set, probably representing the wide range of groundwater residence times and climatic up-concentration factors represented by the constituent waters.

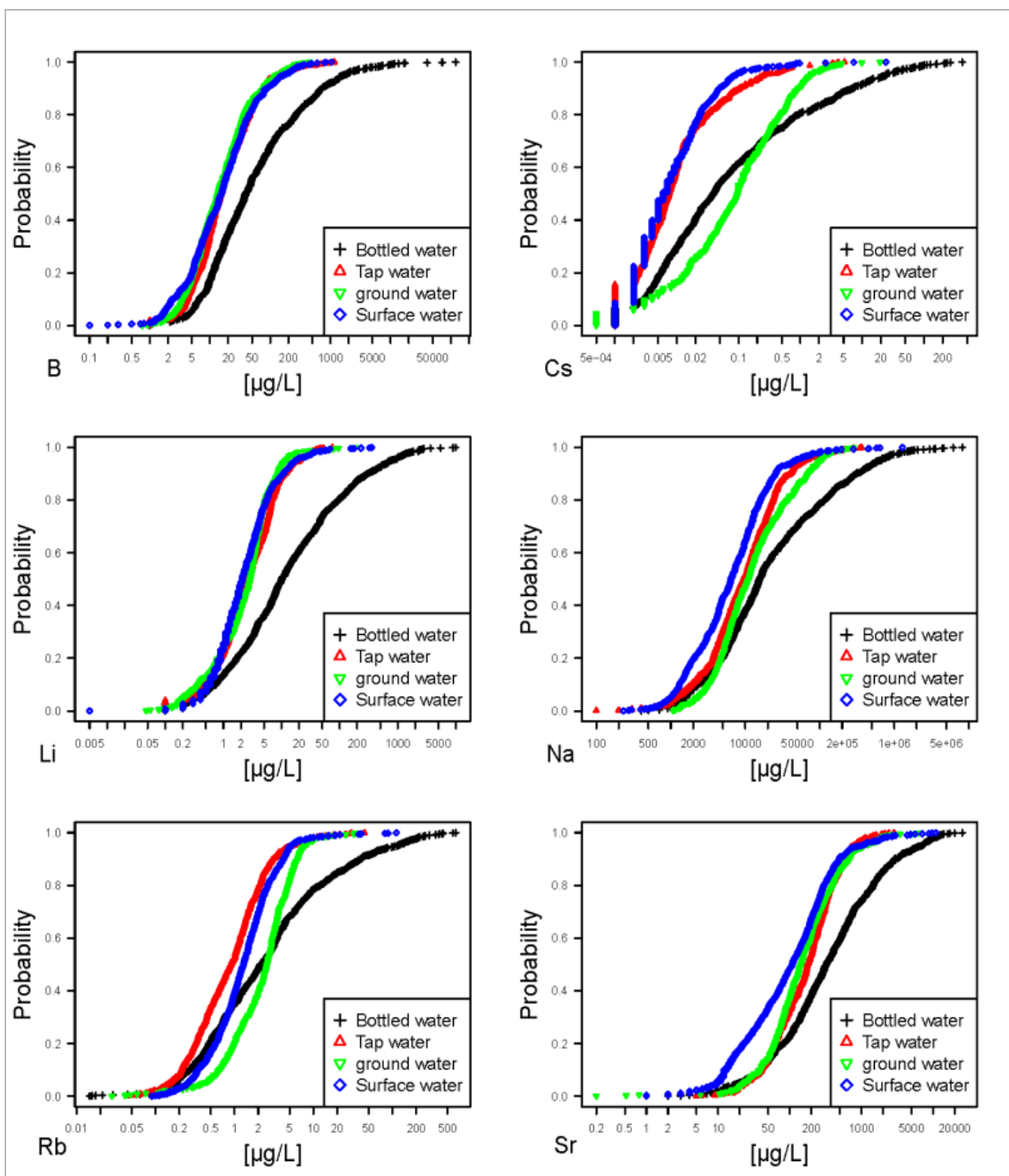


Fig. 3 Empirical cumulative distribution function (ECDF)-plot of B, Cs, Li, Na, Rb and Sr in European bottled water (BW), European tap water (TW), Norwegian hardrock groundwater (GW) and European surface water (SW). Note the logarithmic scale of the x-axis of the plot.

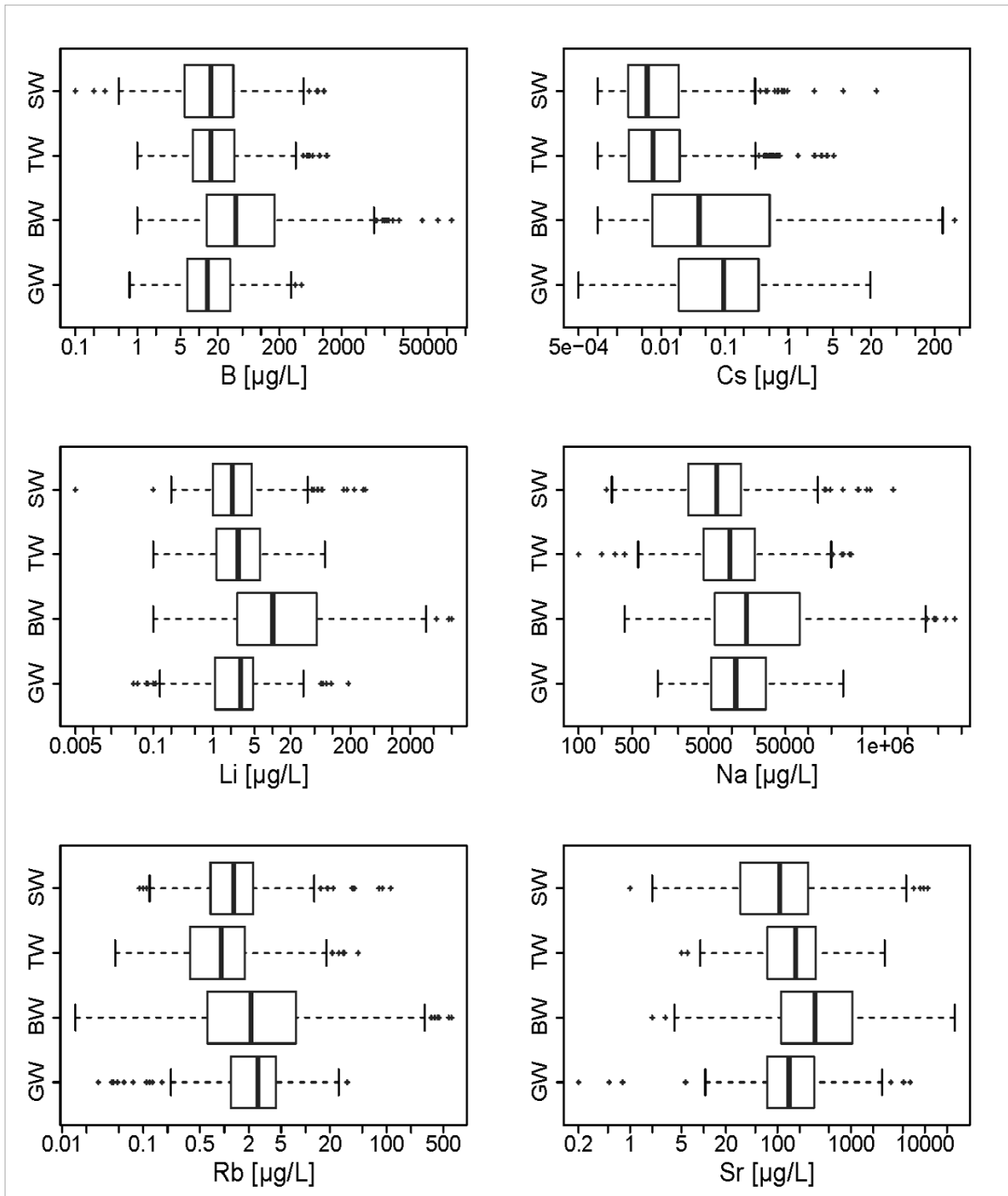


Fig. 4 Boxplots of B, Cs, Li, Na, Rb and Sr, the same suit of elements as in Fig. 3, in European bottled water(BW), European tap water (TW), Norwegian hardrock groundwater (GW) and European surface water (SW). Note the logarithmic scale of the x-axis of the plot.

The elements Al, Ce, Fe, Mn, Th and Ti show very different concentration ranges between the water types in the ECDF-plots (Fig. 5), the distribution function for GW and the SW are strongly shifted to the right (higher concentrations) compared to the curves for BW

(European ground water) and TW. All REE show the same pattern as Ce in Fig. 5. Very specific factors may be responsible for these observations: for example, the generally elevated concentrations of Th and REE in Norwegian hard rock terrain being reflected in the GW data, or the likelihood that elevated Fe and Mn concentrations are deemed aesthetically unacceptable in bottled water (BW) and European tap water (TW) – and are often removed by aeration and/or filtration (see, e.g., Hem, 1985). The high degree of correlation of Fe and Mn in the heatmap of TW (Fig.6) and GW (Fig. S2, supplementary material), and the lack of correlation in SW (Fig. S3, supplementary material) supports this assumption. Additionally, the solubility of many of these elements is strongly dependent on pH and redox conditions, and possibly on complexing with organic carbon (Reimann and Birke, 2010), which may explain the generally higher concentrations in surface waters (organic complexation and lower pH) and Norwegian crystalline bedrock groundwater (often somewhat reducing in nature), compared to tap water and bottled water (often treated); see Fig. 4

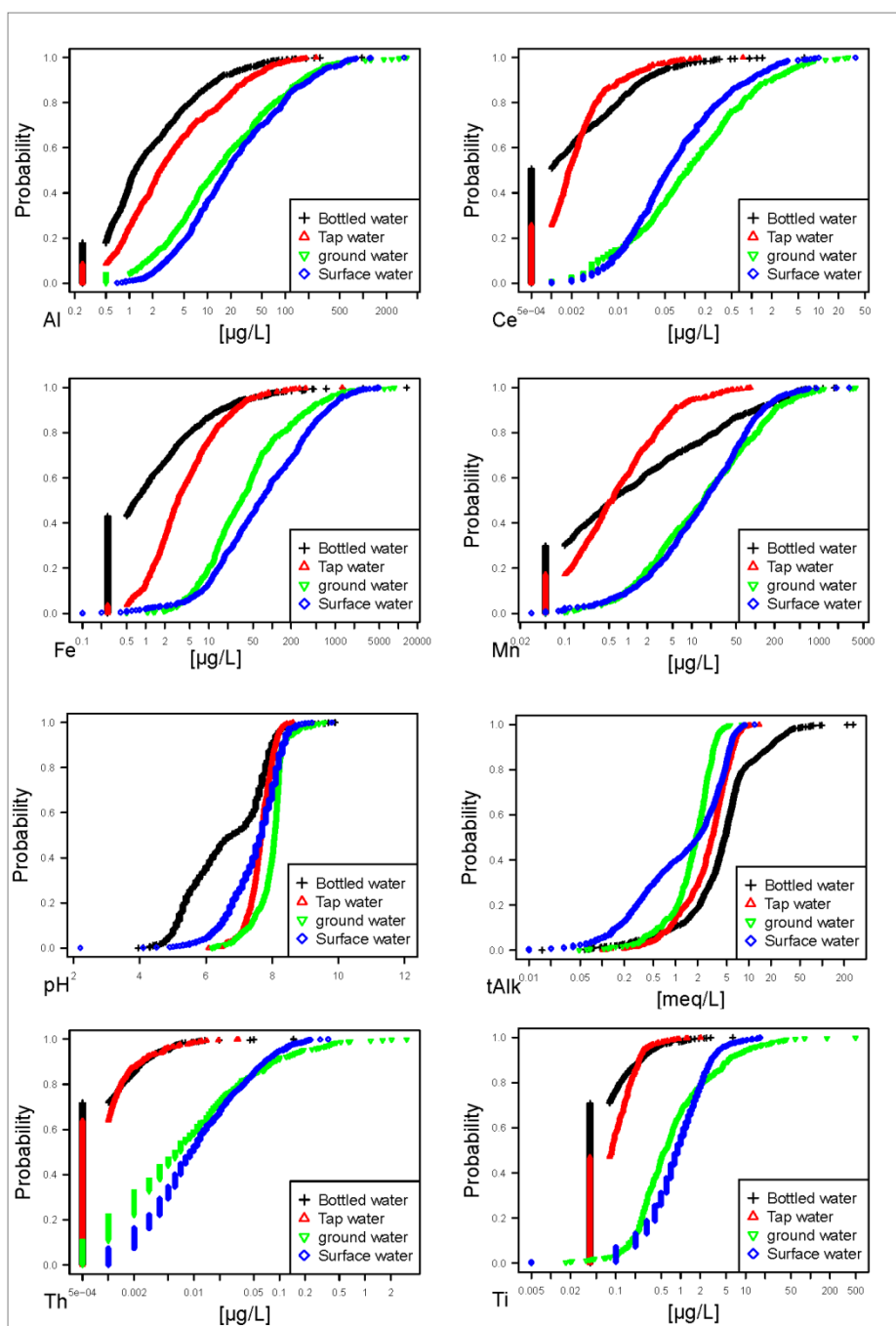


Fig. 5 Empirical cumulative distribution function (ECDF-plot) of Al, Ce, Fe, Mn, pH, tAlk (total alkalinity), Th and Ti in European bottled water (BW), European tap water (TW), Norwegian hardrock groundwater (GW) and European surface water (SW). Note the logarithmic scale of the x-axis of the plot except for pH.

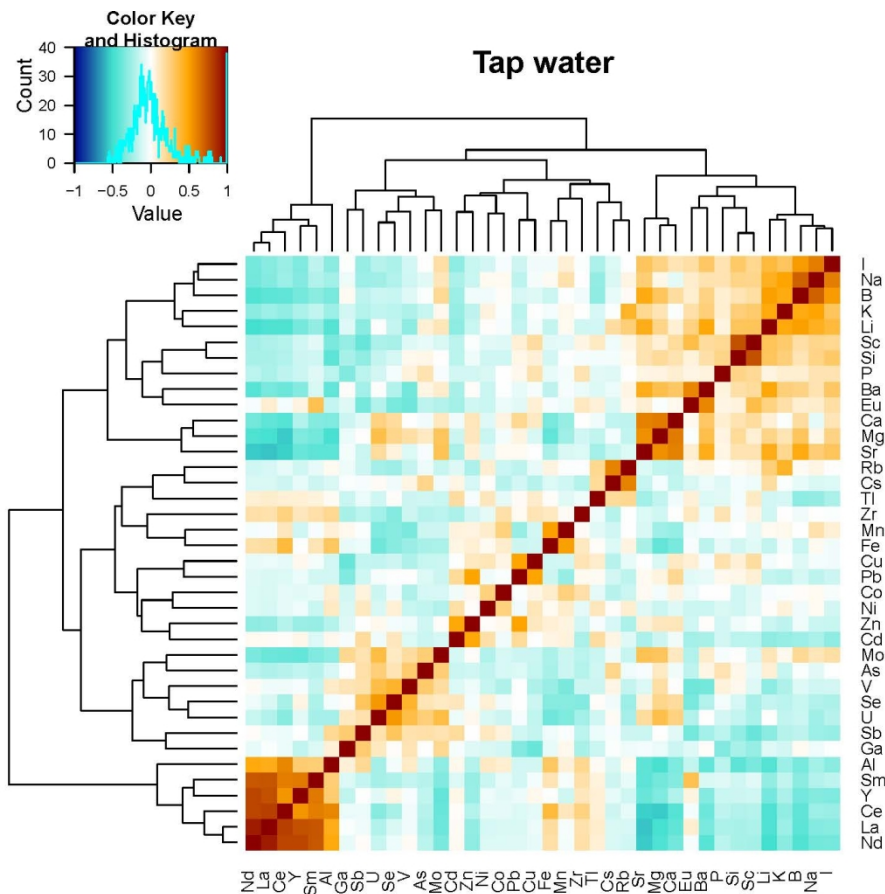


Fig.6 Correlation heatmap for European tap water (TW) (*clr*-transformed data) with elements sorted according to a cluster analysis (see dendrograms). In brief, orange and red colours indicate positive correlations, blue colours indicate negative correlations and the strength of colour indicates the magnitude of correlation coefficient, as shown on the correlation spectrum (inset top left).

### 3.3 Water contamination

Figure 6 shows a correlation heat map for the tap water data set, with only selected elements plotted and those parameters (pH, anions etc.) analysed at other laboratories omitted. The strongest positive correlations can be seen at the upper and lower ends of the map between

- highly soluble elements with few solubility ceilings in natural waters in temperate areas: I, Na, B, K, Li),
- the alkaline earths (Ca, Mg, Sr),

- the REE (at the lower, left side of Fig. 6).

Some areas of weaker correlation exist nearer the centre of the heat map in Fig. 6, and one of these areas of weak positive correlations includes the elements Cd, Cu, Zn, Pb (as well as Mn, Fe, Co and Ni). The concentration of the elements Cd, Cu, Pb and Zn is often assumed to be influenced by anthropogenic factors – i.e., ‘plumbing materials’, pipes, valves, pumps, well casings, solder materials – in the well or distribution network (Zietz et al., 2015), in addition to the geogenic background variation of these elements in the water. In the heatmap of TW (Fig.6) and the heatmaps of GW, SW and BW (Fig.S2-S4, supplementary material) these elements show varying degrees of correlation with each other.

The Norwegian bedrock groundwaters (GW) and the European tap waters (TW) exhibit considerably higher concentrations of Cu and Zn than the bottled waters (BW) or surface waters (SW; Table 1, Fig. 7). Since GW and TW are distributed from well or river intakes, via pipes networks, Banks et al. (2015) and Flem et al. (2015) argued that a substantial portion of the measured concentrations of the elements Cu, Pb and Zn in the European TW dataset may have an anthropogenic (plumbing, pipework, solder) source over most of the concentration range. Also Banks et al. (1998) suggested that the higher concentration of Cu and Zn in the Norwegian ground water might be due to domestic pipe work or water well installations.



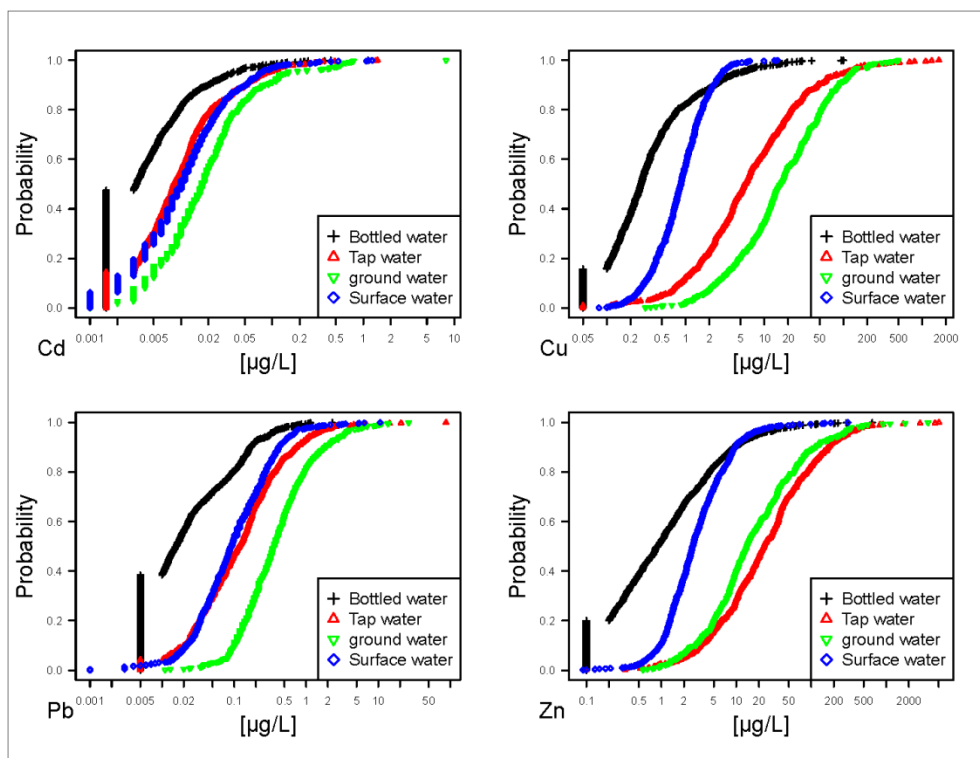


Fig. 7 Empirical cumulative distribution function (ECDF-plot) of Cd, Cu, Pb and Zn in European bottled water (BW), European tap water (TW), Norwegian hardrock groundwater (GW) and European surface water (SW). Note the logarithmic scale of the x-axis of the plot.

### 3.4 Analytical problems

One of the first approaches when studying a new dataset should be to perform a cluster analysis in order to reveal analytical artefacts. This is particular useful when analytical techniques such as ICP-QMS, known to have problems with isobaric interferences, are used. Dendograms in combination with a heatmap provide a powerful tool to discover elemental clustering due to natural and anthropogenic processes, but they can also help to unveil analytical artefacts. The heatmap of TW data (Fig.6) shows, for example, unexpected clustering of Si and Sc and of Eu and Ba. The elemental co-variation of Si and Sc can be seen for all four data sets included in this study (Fig. S2- S4, supplementary material). This observation most probably indicates that the interference correction of  $^{29}\text{Si}^{16}\text{O}$  on  $^{45}\text{Sc}$  is not adequate. This was also suspected by Reimann et al. (2010) and therefore Sc concentrations

in the water samples are not reported in the atlas. Europium, in contrast, would be expected to broadly correlate with the other rare earth elements rather than with Ba. Eu has two isotopes,  $^{151}\text{Eu}$  and  $^{153}\text{Eu}$ , both interfered by Ba ( $^{135}\text{Ba}^{16}\text{O}$  and  $^{137}\text{Ba}^{16}\text{O}$ ) which cannot be avoided by single quadrupole ICP MS instruments, as has been used for the analysis of these waters. Creating scatter plots for all datasets of Ba versus Eu (Fig. 8) reveals that the mathematical interference correction done for Eu has not been adequate for the TW dataset (Birke et al., 2010). It also suggests that Eu in BW seems, at higher concentrations, to be affected by Ba. The other two datasets, Norwegian GW and SW do not show this correlation between Eu and Ba. The whole issue is complicated by the facts that: (i) Eu does, in fact, behave hydrochemically somewhat differently to the other REE (Banks et al., 1999) and (ii) that one might expect to see a “real” covariation between Ba and Eu, as both may conceivably be preferentially mobilised in reducing conditions (Ba, by sulphate reduction removing sulphate from solution and permitting Ba to accumulate without a barite saturation ceiling being reached, and Eu by virtue of the fact that it can exist in a reduced +II oxidation state). If possible, high resolution ICP-MS should be used for the analysis of Eu, or to reveal the reason for the co-variation between Ba and Eu.

Two elements show, compared to the other water types, an unusual bimodal curve in their ECDF plots: Ga in European tap water and Sb in European bottled water (Supplementary material Fig.S1). The less dominant isotope  $^{71}\text{Ga}$  (39.9% abundance) were used for the ICP-QMS analysis of BW and TW. Under some plasma conditions  $^{71}\text{Ga}$  might be interfered by  $^{55}\text{Mn}^{16}\text{O}$ ; however, a scatter plot of Mn and Ga do not reveal any co-variation indicating mass interferences. The overestimation of Ga at low concentrations in TW might be due to crack products from backing pump oil.

The unusual bimodal curve for Sb for bottled water was shown to be due to leaches from both polyethylene terephthalate (PET) bottles and glass bottles (Reimann et al., 2010).

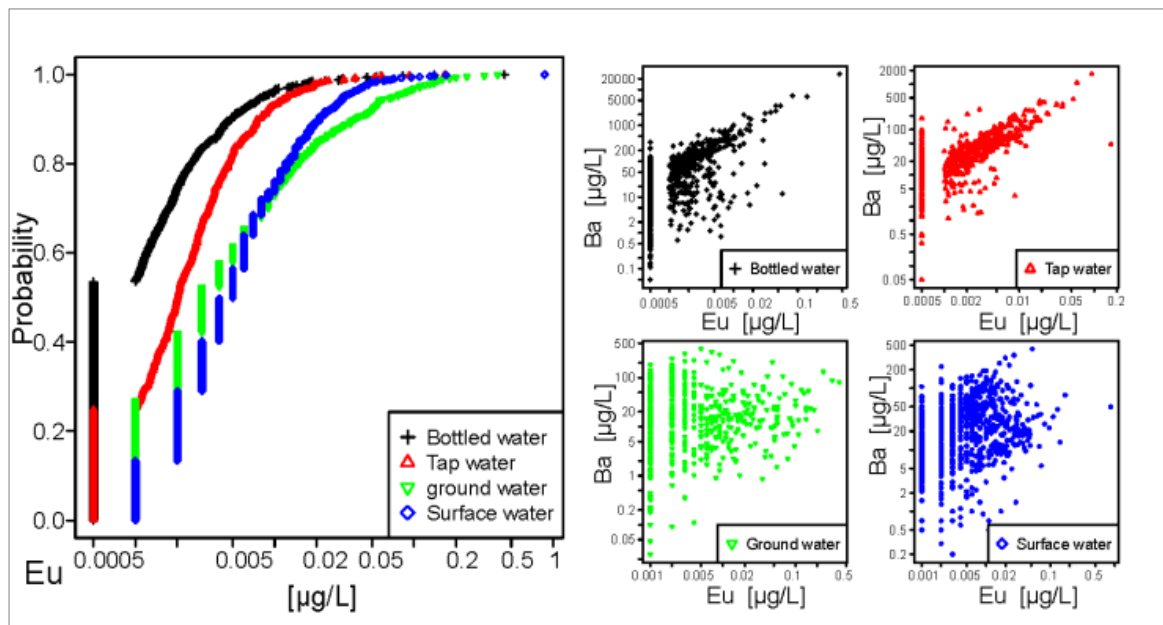


Fig. 8 Empirical cumulative distribution function (ECDF) plot of Eu (left figure) and scatterplot of Ba versus Eu (right figure).

In all the graphics presented here, concentrations below analytical detection limit (DL) have been plotted as  $0.5 \times \text{DL}$ . This produces the 'artefact' of a vertical line at the lowest concentration in the ECDF-plots as can be seen for e.g., Mn and Ti in BW and TW (Fig.5). Many commercial laboratories, as routine, tend to round all numbers to e.g., 3 significant digits (although one can, on request, usually obtain unrounded data from the laboratory). This 'rounding' effect can be observed in, e.g., Th and Ti (Fig. 5) and Eu and Ga (Fig.8) for both SW and GW as apparent vertical lines in both the ECDF-plot and the scatter plots.

### 3.5 Geogenic and natural impact

As has already been noted, several parameters exhibit remarkably similar distributions in different water chemical data sets from across Europe, while others show systematic differences. Some elements, such as As and V, which can be highly mobile (depending on redox and pH conditions), show very similar distributions for BW, TW and GW (Fig. 9). Vanadium's solubility is highest in oxic environments (such as, presumably, surface water), where vanadyl cations ( $\text{VO}^{2+}$  and  $\text{VO}_2^{2+}$ ) predominate. Under more reducing conditions (presumably, deeper groundwater), the less mobile  $\text{V}^{3+}$  dominates, resulting in generally lower concentrations.

The SW data show a median As concentration more than two times higher than BW, TW and Norwegian GW (Table 1, Fig. 9). Arsenic can be fairly mobile in many hydrological environments, but its mobility can be limited due to strong sorption by clays, (especially ferric) hydroxides and organic matter. The composition, grain size distribution and organic matter content of a subsoil or aquifer matrix may thus have a marked influence on measured As-concentrations (Reimann et al., 2003).

A graphical presentation, using boxplots, of the elements shown in ECDF-plots in Fig. 9 is shown in Fig.10. Even though the cumulative distribution curve for surface water is shifted to the right (higher concentrations) for As and V compared to the other waters, the boxplots reveal that 50% of the data (the central box) nearly overlap with the 50% boxes of the other waters.

The distribution of SW is shifted to the right for all the elements shown in Fig. 9 (As, Co, Cr, Ni, Se, V) compared to BW, TW and Norwegian GW.

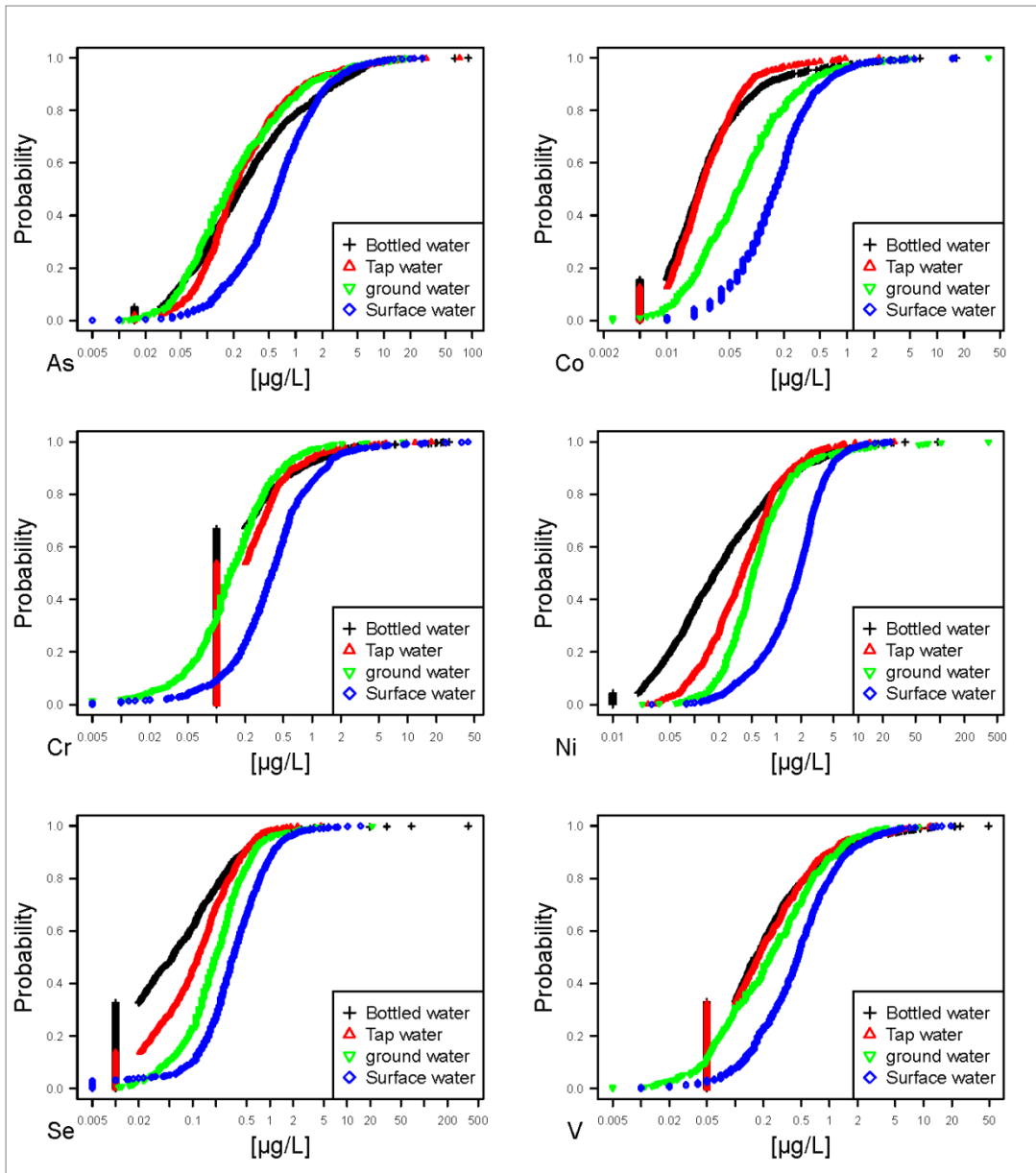


Fig. 9 Empirical cumulative distribution function (ECDF-plot) of As, Co, Cr, Ni, Se and V in European bottled water (BW), European tap water (TW), Norwegian hardrock groundwater (GW) and European surface water (SW).

The influence of the bedrock on soil trace metal composition and the trace elemental composition of water can be assessed by reviewing the spatial distribution and co-variation of specific trace metals that can be linked to the local bedrock. Ni, Cr and Co are all typically enriched in mafic relative to felsic igneous rocks, either as accessory elements in common rock-forming minerals (e.g., olivine, pyroxene, spinel) or in sulphides (e.g., pyrite, sphalerite and pentlandite) and oxides (e.g., chromite; Salminen et al., 2005). Abundances of the three

elements are therefore frequently correlated (Goldschmidt 1954, Rose et al., 1979, Berrow and Reaves, 1986). However, in some sedimentary rocks, such as sulphide-rich shales and their metamorphic equivalents, this correlation may be skewed due to the presence of nickel- (and cobalt-) bearing sulphide minerals and lack of chromite and other chromium-bearing minerals. As can be seen from Fig. 9, the distribution function plots, the Norwegian groundwaters are indeed, relatively enriched in Co and Ni, but not in Cr. In the heatmaps of these waters (Fig.11) the strong correlation of Ni and Co is also shown. The intensity of the orange/red colour in the heatmap indicate the degree of correlation and in this case it indicates that the degree of correlation between Co and Ni in Norwegian hardrock GW is higher than in TW (Fig. 11, top right and bottom left). However, this difference in correlation may also be due to the difference in data quality as detection limit (Tab. 1 and Fig.9).

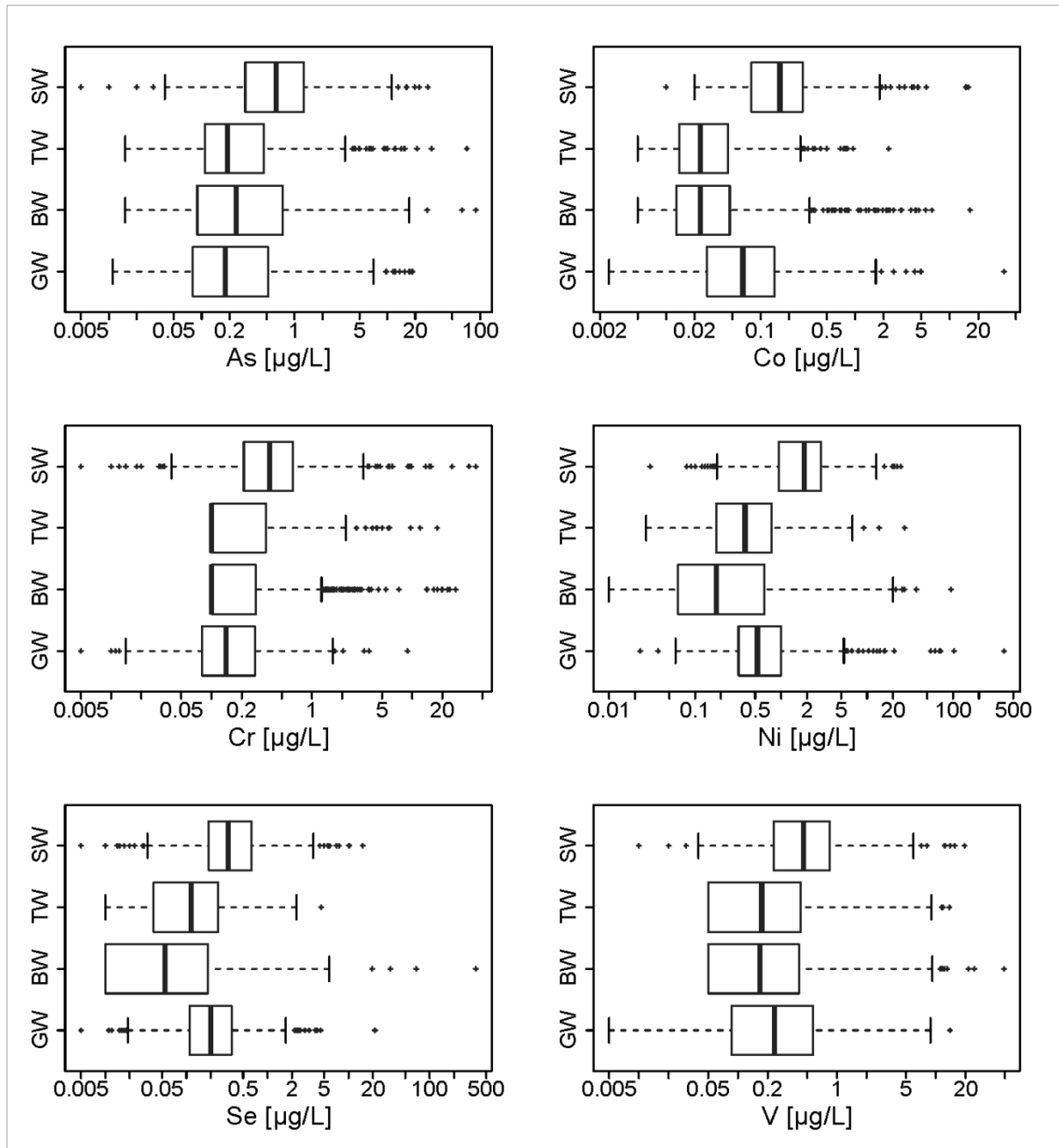


Fig. 10 The same data as in Fig.9 (As, Co, Cr, Ni, Se and V) compared using Turkey boxplots with logarithmic scaling.

Heatmaps in combination with element clustering of the results can help to reveal the unique and the comparable properties of each water dataset. The BW dataset contain 884 samples of groundwater, classified either as natural mineral waters, spring water or bottled drinking water (Reimann and Birke, 2010). This data set shows the generally lowest overall

551 concentration levels for most elements. Only the elements, Al, As, B, Ba, Ca, Co, Cs, Cu, I,  
552 K, La, Li, Mg, Mn, Mo, Na, Nd, Ni, P, Pb, Rb, Sb, Sc, Se, Si, Sr, Tl, U, V, Y, Zn and Zr,  
553 analysed with ICP-MS and ICP-AES (Si) returned more than 60 % of the all analytical results  
554 above detection limit. For multivariate data analysis most data for most elements entered  
555 should be above detection limit, in order to avoid artefacts related to too many data having  
556 the same analytical value. The heatmaps of the four water types in Fig. 11 are therefore based  
557 on the above listed suite of elements. The elements Cu, I, Mn, P, Pb, Sb, Sc, Ta and Zn have  
558 also been excluded due to an anthropogenic influence, unacceptable analytical interference  
559 (Sc) or because some of the elements were not available for all the water data sets (I, P, Tab.  
560 1).



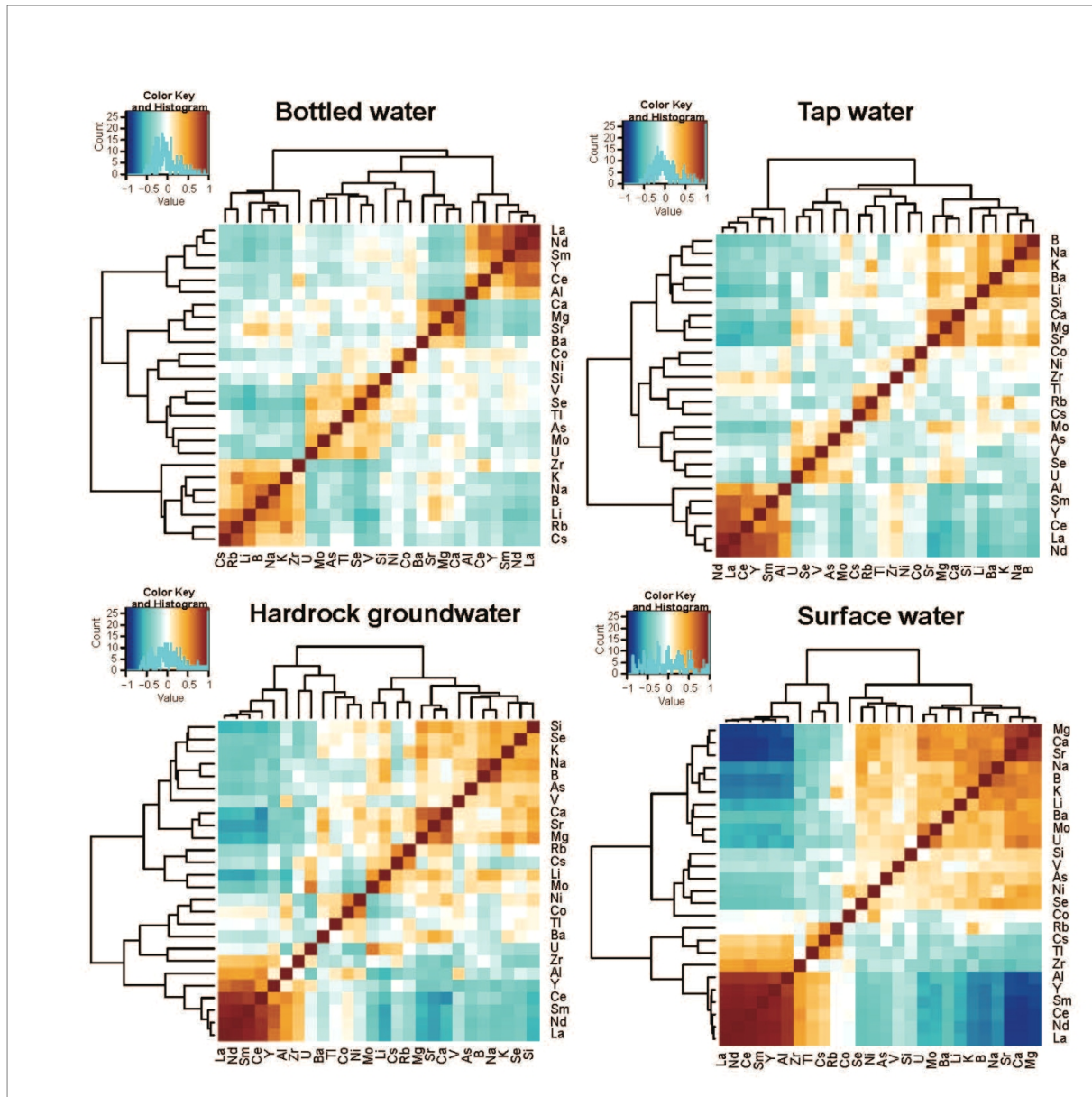


Fig. 11 Correlation analysis for European bottled water (BW), European tap water (TW), Norwegian hardrock groundwater (GW) and European surface water (SW) presented in the form of heat maps. The sequence of the elements is determined by element clustering. Note that the order of elements is different in each map. All data are *clr*-transformed before plotting.

Comparing the four heatmaps in Fig. 11 (note that the order of elements is different in each map) the GW, BW and TW maps exhibit a high degree of common correlations between clusters, compared to the heat map of SW. All the SW samples are collected from running streams in small, second order, drainage basins (Salminen et al., 2005). The samples in both

BW (groundwater) and the Norwegian hardrock GW dataset are affected by different residence time as well as sampling depth. Most of the Norwegian GW is derived from very low production wells with an average depth of less than 80 m, and which may have relatively short residence time (see Misstear et al., 2017). Many of the spring waters and minerals waters (which make up the bottled water BW data set) are required by law to be well-protected against surficial contamination and hydrochemical variability – in practice, this often means relatively deep boreholes tapping groundwater bodies with a long residence time). The TW sample set is even more inhomogeneous because it comprises both surface water and groundwater samples and has in addition been subject to different degrees and types of water treatment. The large difference in cross correlation between SW and the other waters can be seen in the histogram given at the top left of each heatmap. However, despite the inhomogeneity of the sample material, geology is still visible in all the heatmaps.

At the top right in Fig. 11, the heatmap with dendograms for *clr*-transformed data of European BW is shown. The cluster analysis identifies two main clusters where one contains a smaller number of elements Zr, K, Na, B, Li and Rb. This smaller cluster shows a very low degree of correlation with the REE and the sub-cluster containing V, Se, Tl, As, Mo and U.

The 579 samples that constitute the European TW dataset, originate both from groundwater and surface water. In contrast to bottled water, the tap waters chemistry may be significantly affected by more “intensive” water treatment processes (Banks et al., 2015, Flem et al., 2015). The dendrogram divides the parameter suite into two distinct clusters (note the height of the vertical line, indicating the degree of difference between clusters); one small cluster containing the REE together with yttrium and aluminium (which might reflect particulate content or strong pH dependence) and one containing the rest of the suite of elements (which may reflect a variety of progressive water-rock interaction or mineral dissolution processes)

indicating processes caused by water rock interaction. The larger cluster is divided in two sub-clusters where several high soluble elements (that tend to accumulate with residence time and hydrochemical maturity), such as K, Li, B, Na, which correlate strongly with each other, also correlate with other elements e.g., Cs and Rb, in the other sub-clusters.

The relative homogeneity of the SW data set can be observed in the heatmap, showing relatively little cross correlations between the major cluster groups (Fig. 11, bottom right). The elements Zr, Al, Y and REE (which might reflect particulate content or strong pH dependence) form a distinct group that exhibits a low correlation (blue in colour in the heatmap and high vertical lines in the dendogram, Fig. 11) with the alkaline earth elements (which Sr, Mg, Ca, which may have a common mineral origin) or highly soluble elements (Na, K, B, Li), which tend to accumulate with residence time, evapoconcentration and hydrochemical maturity and which may also reflect marine influence.

#### 4. Conclusion

Analytical results for four large water data sets have been compared; (1) bottled water as proxy for European ground water, (2) stream water, (3) tap water and (4) Norwegian hardrock groundwater using ‘simple’ graphical exploratory data analysis tools. The ease of application and power of such apparently ‘simple’ graphical statistical methods do not appear to be sufficiently appreciated by the scientific community. Using a combination of ECDF diagrams and Tukey boxplots, similarities and discrepancies between the four datasets can be detected. A large number of elements show a surprisingly similar concentration range and overall statistical distribution in all water types. This is especially surprising given the different origins of the samples (groundwater and surface water, treated and untreated water). Some elements can be identified as very typical for groundwater (e.g., high concentrations of

Cs, Rb and Na). Examples of contamination from either well installations / distribution network (Cu, Pb, Zn) or from the bottle material (Sb, in the bottled water data set) can also be detected. Some analytical artefacts and isobaric interferences (e.g., Si with Sc) were noted by comparing the results of the four different water types in these diagrams.

Correlation analysis in combination with element clustering of the results provides a further powerful graphical data analysis tool, which can be used to generate hypotheses concerning some of the key processes influencing the four water types.

When working with CoDa, it turns out that there is a clear difference whether absolute concentrations or *clr* or *ilr* transformed data are to be studied. Both approaches are acceptable but deliver different information. There exist many instances where absolute concentration data are to be preferred. In these cases, however, the researcher should be aware that he/she is not working with the usual Euclidean geometry on which standard statistical methods rely. Great care is needed in terms of which statistical methods can be applied; for example, already the standard deviation loses its meaning when working with CoDa. One should thus rely on relatively simple statistical (non-parametric) techniques. In this paper, the use of rank statistics, the ECDF and boxplots is demonstrated. In multivariate analysis, the focus shifts towards on the proportionality between the elements and the data should be transformed into the correct geometry for being able to use techniques based on Euclidean distances via an appropriate log-ratio transformation. This applies even to the relatively commonly-used correlation analyses. Working with the correct geometric space quite often results in a more demanding interpretation of the results.

Acknowledgements

The EGG Project Team is thanked for the bottle water and tap water survey and the FOREGS Project Team for the surface water survey. We highly appreciate that the datasets are made freely available to the scientific society. Bjørn Frengstad (NGU) is thanked for providing the Norwegian hardrock groundwater data set.

The authors would like to thank the two reviewers for their in-depth, very insightful and in parts quite demanding reviews which helped to improve the clarity of the article.

## References

Aitchison, J., 1986. The statistical analysis of compositional data (Monographs on statistics and applied probability). London: Chapman and Hall.

An, Tsujimura, Le Phu, Kawachi, & Ha. (2014). Chemical Characteristics of Surface Water and Groundwater in Coastal Watershed, Mekong Delta, Vietnam. *Procedia Environmental Sciences*, 20, 712-721

Banks, D., Midtgard, A.K., Frengstad, B., Krog, J.R., Strand, T.R., 1998. The chemistry of Norwegian groundwaters: II. The chemistry of 72 groundwaters from quaternary sedimentary aquifers. *Science of the Total Environment* 222, 93-105. doi: 10.1016/S0048-9697(98)00291-5

Banks, D., Hall, G., Reimann, C., Siewers, U., 1999. Distribution of rare earth elements in crystalline bedrock groundwaters: Oslo and Bergen regions, Norway. *Applied Geochemistry* 14(1), 27-39. doi: 10.1016/S0883-2927(98)00037-7

Banks, D., 2014. A Hydrogeological Atlas of Faryab Province, Northern Afghanistan. Norplan report for Afghan Ministry of Rural Rehabilitation and Development, funded by NORAD. Published November 2014 by Asplan VIAK AS, Kristiansand, Norway and available at [www.norplan.af](http://www.norplan.af). doi: 10.13140/RG.2.1.1528.9444

Banks, D., Birke, M., Flem, B., Reimann, C., 2015. Inorganic chemical quality of European tap-water: 1. distribution of parameters and regulatory compliance. *Applied Geochemistry*, 59, 200-210. doi: 10.1016/j.apgeochem.2014.10.016

Bates, D., Maechler, M., 2016. Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.2-6. <https://CRAN.R-project.org/package=Matrix>

Berrow, M.L., Reaves, G.A., 1986. Total chromium and nickel contents of Scottish soils. *Geoderma* 37, 1, pp 15-27.

Bertoldi, D., Bontempo, L., Larcher, R., Nicolini, G., Voerkelius, S., Lorenz, G.D., Ueckermann, H., Froeschl, H., Baxter, M.J., Hoogewerff, J., Brereton, P., 2011. Survey of the chemical composition of 571 European bottled mineral waters. *Journal of Food Composition and Analysis* 24, 376-385.

Bhowmik, A.K., Alamdar, A., Katsoyiannis, I., Shen, H., Ali, N., Ali, S.M., Bokhari, H., Schafer, R.B., Eqani, S., 2015. Mapping human health risks from exposure to trace metal contamination of drinking water sources in Pakistan. *Science of the Total Environment* 538, 306-316.

Birke, M., Reimann, C., Demetriades, A., Rauch, U., Lorenz, H., Harazim, B., et al., 2010. Determination of major and trace elements in European bottled mineral water - Analytical methods. *Journal of Geochemical Exploration*; 107: 217-226.

694 Buccianti, A., Pawlowsky-Glahn, V., 2005. New perspectives on water chemistry and  
695 compositional data analysis. *Mathematical Geology* 37(7), 703-727. doi: 10.1007/s11004-  
696 005-7376-6

697 Cao, Y.J., Tang, C.Y., Song, X.F., Liu, C.M., Zhang, Y.H., 2016. Identifying the  
698 hydrochemical characteristics of rivers and groundwater by multivariate statistical  
699 analysis in the Sanjiang Plain, China. *Applied Water Science* 6, 169-178.

700 Dragon, K., Marciniak, M., 2010. Chemical composition of groundwater and surface water in  
701 the Arctic environment (Petuniabukta region, central Spitsbergen), *Journal of Hydrology*,  
702 386(1), 160-172.

703 Dragon, K., Gorski, J., 2015. Identification of groundwater chemistry origins in a regional  
704 aquifer system (Wielkopolska region, Poland). *Environmental Earth Sciences*, 73(5),  
705 2153-2167.

706 Egozcue, J.J., Pawlowsky, Glahn, V., Mateu-Figueraz, G., Barcelo-Vidal, C., 2003. Isometric  
707 logratio transformations for compositional data analysis. *Math. Geol.* 35:279-300

708 Engle, M.A., Gallo, M., Schroeder, K.T., Geboy, N.J., Zupancic, J.W., 2014. Three-way  
709 compositional analysis of water quality monitoring data. *Environmental and Ecological*  
710 *Statistics*, 21, 565-581. doi: 10.1007/s10651-013-0268-x.

711 Engle, M.A., Rowan, E.L., 2014. Geochemical evolution of produced waters from hydraulic  
712 fracturing of the Marcellus Shale, northern Appalachian Basin: A multivariate  
713 compositional data analysis approach. *International Journal of Coal Geology* 126, 45-56.  
714 doi: 10.1016/j.coal.2013.11.010

715 Filzmoser, P., (2015). StatDA: Statistical Analysis for Environmental Data. R package  
716 version 1.6.9. <https://CRAN.R-project.org/package=StatDA>

717 Filzmoser, P., Hron, K., Reimann, R., 2010. The bivariate statistical analysis of  
718 environmental (compositional) data. *Science of the Total Environment*, 408(19), 4230-  
719 4238.

720 Filzmoser, P., Hron, K., Reimann, C., 2009a. Principal component analysis for compositional  
721 data with outliers. *Environmetrics* 20: 621-632.

722 Filzmoser, P., Hron, K., Reimann, C., Garrett, R., 2009b. Robust factor analysis for  
723 compositional data. *Computers & Geosciences* 35, 1854–1861.

724 Filzmoser, P., Hron, K., Reimann, C., 2009c. Univariate statistical analysis of environmental  
725 (compositional) data: Problems and possibilities. *Science of the Total Environment*, Vol. 407, pp.  
726 6100-6108.

727 Flem, B., Reimann, C., Birke, M., Banks, D., Filzmoser, P., Frengstad, B., 2015. Inorganic  
728 chemical quality of European tap-water: 2. Geographical distribution. *Applied*  
729 *Geochemistry*, 59, 211–224. doi: 10.1016/j.apgeochem.2015.01.016

730 Frengstad, B., Skrede, A.K.M., Banks, D., Krog, J.R., Siewers, U., 2000. The chemistry of  
731 Norwegian groundwaters: III. The distribution of trace elements in 476 crystalline  
732 bedrock groundwaters, as analysed by ICP-MS techniques. *Science of the Total Environ.*  
733 246, 21-40. doi: 10.1016/S0048-9697(99)00413-1

734 Frengstad, B., Banks, D., Siewers, U., 2001. The chemistry of Norwegian groundwaters: IV.  
735 The pH-dependence of element concentrations in crystalline bedrock groundwaters.  
736 *The Science of the Total Environment*, 277, 101-117. doi: 10.1016/S0048-  
737 9697(00)00867-6

738 Goldschmidt, V.M., 1954, *Geochemistry*. Ed. by Alex Muir, Clarendon Press, 1954

739 Hem, J. D., 1985. *Study and Interpretation of the Chemical Characteristics of Natural Water*,  
740 3rd ed. Alexandria, VA: Department of the Interior, U.S. Geological Survey, Water-  
741 Supply Paper 2254.

742 Kynclova, P., Hron, K., and Filzmoser, P., 2017. Correlation Between Compositional Parts  
743 Based on Symmetric Balances. *Mathematical Geosciences*, 49(6), 777-796.

744 Luís, A. T., Teixeira, P., Almeida, S. F. P., Matos, J. X., Da Silva, E. F., 2011. Environmental  
745 impact of mining activities in the Lousal area (Portugal): Chemical and diatom  
746 characterization of metal-contaminated stream sediments and surface water of Corona  
747 stream. *Science of the Total Environment*, 409(20), 4312-4325.

748 Misstear, B., Banks, D., Clark, L. 2017. *Water wells and boreholes*, 2<sup>nd</sup> edition. Wiley,  
749 Chichester (page 366, Text Box 8.1 ‘Groundwater chemistry as a guide to vulnerability’)

750 O Dochartaigh, B.E., MacDonald, A.M., Fitzsimons, V., Ward, R., 2015. Scotland's aquifers  
751 and groundwater bodies. Nottingham, UK, British Geological Survey, 63pp. (OR/15/028)  
752 (Unpublished)

753 Pawlowsky-Glahn, V., Buccianti, A., 2011. *Compositional data analysis: Theory and*  
754 *applications*. Wiley, Chichester, 400 pp.

755 R Core Team., 2016. *R: A language and environment for statistical computing*. R Foundation  
756 for statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

757 Reimann, C., Filzmoser, P., Hron, K., Kynčlová, P., Garrett, R. G., 2017. A new method for  
758 correlation analysis of compositional (environmental) data – a worked example, *Science*  
759 *of The Total Environment*, 607, 965-971.

760 Reimann; C., Siewers; U., Tarvainen; T., Bityukova; L., Eriksson; J., Giucis; A.,  
761 Gregorauskiene; V., Lukashev; V.K., Matinian; N.N., Pasieczna, A., 2003. *Agricultural*  
762 *Soils in Northern Europe: A Geochemical Atlas*. Schweizerbart Science Publishers,  
763 Stuttgart, Germany. 279 pp

764 Reimann, C., Filzmoser, P., Garrett, R. G., Dutter, R., 2008. *Statistical Data Analysis*  
765 *Explained: Applied Enviromental Statistics with R*. John Wiley & Sons, Ltd, Chichester  
766 (UK).

767 Reimann, C., Birke, M. (eds.), 2010. *Geochemistry of European Bottled Water*. Borntraeger /  
768 Schweizerbart, Stuttgart, Germany. 268 pp. ISBN 978-3-443-01067-6.

769 Reimann, C., Birke, M., Filzmoser, P., 2010. Bottled drinking water: Water contamination  
770 from bottle materials (glass, hard PET, soft PET), the influence of colour and  
771 acidification. *Applied Geochemistry* 25, 1030-1046.

772 Rose, A.W., Hawkes, H.E., Webb, J.S., 1979. *Geochemistry in mineral exploration-2nd ed.*,  
773 Academic Press inc., London, Ltd. 657pp

774 Salminen, R. (Chief-editor), Batista, M. J., Bidovec, M., Demetriades, A., De Vivo, B., De  
775 Vos, W., Duris, M., Gilucis, A., Gregorauskiene, V., Halamic, J., Heitzmann, P., Lima,  
776 A., Jordan, G., Klaver, G., Klein, P., Lis, J., Locutura, J., Marsina, K., Mazreku, A.,  
777 O'Connor, P. J., Olsson, S.Å., Ottesen, R.-T., Petersell, V., Plant, J.A., Reeder, S.,  
778 Salpeteur, I., Sandström, H., Siewers, U., Steenfelt, A., Tarvainen, T., 2005. *Geochemical*  
779 *Atlas of Europe. Part 1: Background Information, Methodology and Maps*. Espoo,  
780 Geological Survey of Finland, 526 pages, 36 figures, 362 maps.



781 Shand, P., Edmunds, W.M., Lawrence, A.R., Smedley, P., Burke, S., 2007. The natural  
782 (baseline) quality of groundwater in England and Wales. Environment Agency, 72pp.  
783 (RR/07/006)

784 Sinclair, A., 1974. Selection of threshold values in geochemical data using probability  
785 graphs. *Journal of Geochemical Exploration*, 3(2), 129-149.

786 Sinclair, A., 1991. A fundamental approach to threshold estimation in exploration  
787 geochemistry: Probability plots revisited. *Journal of Geochemical Exploration*, 41(1), 1-  
788 22.

789 Stanley, C. R., and Sinclair, A. J., 1987. Anomaly recognition for multi-element geochemical  
790 data - A background characterization approach. *Journal of Geochemical Exploration*,  
791 29(1-3), 333-353.

792 Templ, M., Hron, K., Filzmoser, P., 2011. robCompositions: an R-package for robust  
793 statistical analysis of compositional data. In V. Pawlowsky-Glahn and A. Buccianti,  
794 editors, *Compositional Data Analysis. Theory and Applications*, pp. 341-355, John Wiley  
795 & Sons, Chichester (UK).

796 Tennant, C. B., White, M. L., 1959. Economic Geology and the Bulletin of the Society of  
797 Economic Geologists, 54(7), 1281-1290.

798 Tukey, J., 1977. *Exploratory data analysis* (Addison-Wesley series in behavioral science).  
799 Reading, Mass: Addison-Wesley.

800 Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W.H.A., Lumley, T.,  
801 Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B., 2016. gplots:  
802 Various R Programming Tools for Plotting Data. R package version 3.0.1.  
803 <https://CRAN.R-project.org/package=gplots>.

804 Wang, M.Q., Ren, T.X., Liu, Y.H., Liu, S.X., 1995. Stream water, an ideal medium for  
805 rapidly locating polymetallic mineralization in forest swamp terrain: A case study in the  
806 Da Hinggan Mts, northern Heilongjiang, China. *Journal of Geochemical Exploration* 55,  
807 257-264.

808 Zietz, B.P., Richter, K., Lass, J., Suchenwirth, R., Huppmann, R., 2015. Release of Metals  
809 from Different Sections of Domestic Drinking Water Installations. *Water Quality*  
810 *Exposure and Health* 7, 193-204.